Assessing Reading Comprehension in Bilinguals

Assessing Reading Comprehension in Bilinguals

Diane August
Center for Applied Linguistics, Washington, DC

David J. Francis
University of Houston

Han-Ya Annie Hsu
Harvard University

Catherine E. Snow
Harvard University

The Elementary School Journal
Volume 107, Number 2

0013-5984/2006/10702-0006$05.00

Abstract

A new measure of reading comprehension, the Diagnostic Assessment of Reading Comprehension (DARC), designed to reflect central comprehension processes while minimizing decoding and language demands, was pilot tested. We conducted three pilot studies to assess the DARC's feasibility, reliability, comparability across Spanish and English, developmental sensitivity, and relation to standardized measures. The first study, carried out with 16 second- through sixth-grade English language learners, showed that the DARC items were at the appropriate reading level. The second pilot study, with 28 native Spanish-speaking fourth graders who had scored poorly on the Woodcock-Johnson Language Proficiency Reading Passages subtest, revealed a range of scores on the DARC, that yes-no answers were valid indicators of respondents' thinking, and that the Spanish and English versions of the DARC were comparable. The third study, carried out with 521 Spanish-speaking students in kindergarten through grade 3, confirmed that different comprehension processes assessed by the DARC (text memory, text inferencing, background knowledge, and knowledge integration) could be measured independently, and that DARC scores were less strongly related to word reading than Woodcock-Johnson comprehension scores. By minimizing the need for high levels of English oral proficiency or decoding ability, the DARC has the potential to reflect the central comprehension processes of second-language readers of English more effectively than other measures.

The purpose of this article is to consider the challenges of assessing comprehension in second-language (L2) readers and to report on three studies conducted to develop and validate a new measure of reading comprehension called the Diagnostic Assessment of Reading Comprehension (DARC). The DARC, based on the assessment first de-

vised by Potts and Peterson (1985) and extended by Hannon and Daneman (2001), was designed to assess students' performance on four central comprehension processes: remembering newly read text, making inferences licensed by the text, accessing relevant background knowledge, and making inferences that require integrating background knowledge with the text. By minimizing the need for high levels of English oral proficiency or decoding ability, the DARC has the potential to reflect the comprehension skills of L2 readers of English.

Multiple Determinants of Success in Reading Comprehension

Successful reading comprehension reflects the presence of many component capabilities. Comprehension relies on decoding skills (reading words accurately and fluently, accessing lexical representations), knowledge in several domains (vocabulary, linguistic structure, and discourse as well as world knowledge), and cognitive processing capacities (memory for text, accessing relevant background knowledge, drawing justified inferences). Because successful comprehension requires inputs from all these domains of knowledge and processing, it can be disrupted by a failure in any of them, even if the reader is competent in the other ones. Comprehension is like a chemical reaction, which can be constrained by too little of any one of the elements necessary in the reaction, even if the others are present in abundant quantities. This limiting-element conceptualization of comprehension helps explain why comprehension is so vulnerable; breakdown of comprehension can be caused by failures of word-reading automaticity, of familiarity with key vocabulary words in the text, of background knowledge presupposed by the text, of knowledge of discourse features used in the text, of interest in the topic, of inferencing, or of formulating or recognizing a purpose for reading the text. A limitation in a single domain may generate poor comprehension of a particular text even among readers with generally strong comprehension skills. Disruption of comprehension by a single limitation in the face of generally good comprehension skills is, unfortunately, invisible in standardized comprehension assessments, which can produce low scores for readers who would score high if one or two characteristics of the text or the situation were changed.

Even more important, there is little basis for deciding the relative importance of these factors in determining poor comprehension outcomes for individual children or groups of children. Knowing how each factor contributes to comprehension could help in designing optimally differentiated comprehension instruction. In other words, if a group of children comprehends poorly because of failure to draw appropriate inferences, then attention to strategies for constructing inferences makes instructional sense. However, if children show normal ability to form inferences but lack relevant vocabulary knowledge, then focusing instruction on forming inferences is a waste of time. The ultimate purpose of the DARC is to provide teachers with a better basis for adapting instruction to individual students' needs by helping to identify subgroups of struggling comprehenders.

Reasons Children Fail at Comprehension

Many children score poorly on comprehension assessments because their word reading is inaccurate (e.g., Adams, 1990; Gough & Tunmer, 1986; Vellutino, 1979, 1987); children growing up in poverty often fall into this group (National Research Council, 1998). Some children who read accurately, though, fail at comprehension because of inadequate reading fluency (e.g., Perfetti, 1985); children with little access to print or those experiencing instruction that does not emphasize regular reading may be overrepresented in this group (e.g., Stanovich, 1991). Children with little interest in reading also are likely to show poor comprehension (e.g., Guthrie, Wigfield, & VonSecker,

2000; Sweet, Gughrie, & Ng, 1998), perhaps in part because of their restricted opportunities for practice. Still other children with accurate and fluent word-reading skills fail at comprehension because of poor vocabulary and/or limited background knowledge (e.g., Bradley & Bryant, 1983; Hulme, Muter, Snowling, & Stevenson, 2004); children growing up in poverty (Hart & Risley, 1995) and second-language speakers (National Research Council, 1997) show heightened risk of falling into this group. Evidence that explicit strategy instruction improves comprehension (National Reading Panel, 2000) suggests that comprehension is difficult for children who lack techniques for self-monitoring or self-correction. The RAND Reading Study Group (2002) suggested that such failures might be related to the absence of conscious or self-initiated purposes for reading.

The particular challenges of reading comprehension for children from low-income families and for English-language learners (ELLs) deserve mention. Such children typically have smaller vocabularies, less background knowledge relevant to the texts they encounter, and less familiarity with mainstream discourse patterns than high-socioeconomic-status or English-only readers, though there is no a priori reason to assume they are more likely to have difficulties with forming inferences. ELLs may find text memory, an important correlate of scholastic achievement (Gathercole & Pickering, 2000), a particular challenge in English. Children for whom initial reading instruction was a lengthy and/or frustrating process are also likely to develop reduced motivation to read and limited interest in school-assigned reading materials (National Research Council, 1998). Assessing the key limiting factors for different children is crucial to designing effective, targeted instruction for all of them.

The Need for Better Comprehension Assessments

Sorting out optimal instruction for every learner requires having information about which aspects of reading are causing any child's comprehension breakdown. Current comprehension assessments provide limited help with this task. First, these assessments are atheoretical in design. Standardized comprehension assessments are generally "portmanteau measures"—a single score reflects a large domain. Thus, they do not reflect the many factors that influence outcomes.

Second, existing comprehension assessments identify poor readers but do not isolate the determinants of poor performance. Readers with poor skills across the board cannot be distinguished from readers whose comprehension outcomes are limited only by background knowledge, fluency, or another specific factor. Thus, teachers have little guidance from test results concerning what child skills they should focus on.

Furthermore, the most helpful assessments provide information about children's strengths as well as their weaknesses. For example, readers might be very good at memory for new information presented in text and at drawing inferences—strengths that teachers could build on—but be unable to display those capabilities if too many words in the text are unfamiliar. Current assessments are particularly unhelpful in providing information about the comprehension processing of readers with poor vocabularies, for example, ELLs who may bring strong inferencing skills and good strategy use to reading but who are still in the early stages of English vocabulary learning.

A Model for More Diagnostic Comprehension Assessments

Measures of central comprehension processes minimally influenced by other factors have been previously developed, for use with both ELL and English monolingual populations, by Potts and Peterson (1985) and Hannon and Daneman (2001). Potts and Petersen (1985) developed a test that isolated four processes that occur during successful reading comprehension

(Dixon, LeFevre, & Twilley, 1998; Engle, Nations, & Cantor, 1990; Haengi & Perfetti, 1994; Palmer, MacLeod, Hunt, & Davidson, 1985): (*a*) recalling from memory new information presented in the text, which we call *text memory*; (*b*) making novel inferences based on information provided in the text, *text inferencing*; (*c*) accessing relevant prior knowledge from long-term memory, *knowledge access*; and (*d*) integrating accessed prior knowledge with new text information, *knowledge integration*. Scores on the Potts and Peterson (1985) test predicted performance on a general measure of reading comprehension, and scores reflecting the four components related to other, independent tests of those components. In their assessment, reading passages consisted of three sentences that described relations among a set of real and artificial terms (e.g., *a jal is larger than a toc, a toc is larger than a pony, and a beaver is larger than a caz*). Using the information in the text and world knowledge, participants could construct a five-item linear ordering (jal > toc > pony > beaver > caz). Participants read and studied the paragraph and then responded to true-false statements of four types. Text memory statements (e.g., *a jal is larger than a toc*) tested information explicitly mentioned in the paragraph. Text inferencing statements (e.g., *a jal is larger than a pony*) required integrating information across propositions in the text (i.e., *a jal is larger than a toc; a toc is larger than a pony*); no prior knowledge was required. Knowledge access statements (e.g., *a pony is larger than a beaver*) could be answered by accessing prior knowledge; no information from the text was required. Knowledge integration statements (e.g., *a toc is larger than a beaver*) required integrating prior knowledge (ponies are larger than beavers) with a text-based fact (i.e., *a toc is larger than a pony*).

Potts and Peterson found that text memory and text inferencing were highly correlated with each other, but neither was correlated with knowledge access. Knowledge integration was correlated with the two text-based components as well as with the pure prior knowledge component. These correlations suggest that the ability to remember new information and the tendency to use world knowledge are separable. Hannon and Daneman (2001) developed a version of the test with more complex texts, for use with university students. The Hannon and Daneman test also proved to be a valid measure of the four components of reading comprehension processing (Dixon et al., 1988; Engle et al., 1990; Haengi & Perfetti, 1994; Palmer et al., 1985), as shown by correlations of the four scores with performance on comprehension tests designed to assess components of reading comprehension. Furthermore, the Hannon and Daneman test was brief and easy to administer and accounted for a substantial proportion of the variance in performance on a global, standardized test of reading comprehension (the Nelson-Denny test). Scores on knowledge integration were the best predictor of Nelson-Denny scores. The text and test designs Potts and Peterson and Hannon and Daneman used formed the basis for pilot work reported here.

Purpose of the Study

Building on previous work, we developed the Diagnostic Assessment of Reading Comprehension (DARC) for use with ELLs in the elementary grades. Because of our focus on ELL readers, we needed texts that were even simpler than those Potts and Petersen (1985) used. The DARC uses simple and highly decodable words and severely restricts the need for background knowledge in texts that require sophisticated inferencing and knowledge integration.

In this article we present data from three pilot studies designed to refine and validate the measure. The purpose of the first pilot study, carried out in Washington, DC, was to determine that the DARC items were at the appropriate reading level for elementary students of limited English proficiency. The purpose of the second study, carried out in Chicago and Boston, was to provide

an initial validation of the DARC by determining whether Spanish-speaking ELLs who had scored very poorly on a standardized comprehension assessment would show a range of scores on the DARC, as well as to assess the validity of participants' yes-no responses and to compare performance on Spanish and English versions of the DARC. The third pilot study, carried out in Texas, was conducted to estimate developmental sensitivity, reliability, and validity of the DARC subscales using a larger sample.

Method

Participants

The first sample consisted of 16 English-language learners. Two to four students from each of grades 2 through 6 were selected by teachers to represent students with differing levels of English proficiency. All children in this sample had some proficiency in English but were nevertheless English-language learners. All but one of the children spoke Spanish as a first language. The children were participating in a dual-immersion bilingual program in Washington, DC. In the second sample, subjects were 28 native Spanish speakers who were currently in all-English instruction in fourth grade, though all of them had received initial literacy instruction in Spanish, in schools in Chicago ($n = 15$) and Boston ($n = 13$). All were part of a larger longitudinal study of transfer of literacy skills from Spanish to English. The students selected for testing with the DARC had scored in the lowest third of the 168 students in the larger Boston and Chicago sample on the Woodcock Language Proficiency Battery reading comprehension subtest. The third sample consisted of 521 Spanish-speaking students in kindergarten through grade 3 living in Houston and Brownsville, Texas.

Measures

We started with the Potts and Peterson materials, which included only one relational feature in their texts. We made additional adjustments for younger children and ELLs: (*a*) with the Lexile Framework as a guide (http://www.lexile.com), writing texts at the second-grade reading level to ensure that most children could read them, (*b*) using vocabulary that young children were likely to know, (*c*) employing very simple syntactic structures, and (*d*) embedding the relational propositions (*A culp is faster than a cat*) in a more narrative text (e.g., *Mary has four pets*) to provide a familiar genre and more context. Thus, our text minimized the effect of differences across children in decoding skills, vocabulary, and linguistic sophistication—abilities that may mask skills at the heart of reading comprehension (memory for text and making text-based and integrative inferences). We thought this task might be especially useful in revealing inferencing skills of ELL children with normal comprehension processes but limited English proficiency.

The resulting assessment consisted of a brief passage of narrative text that described relations among five entities, where three of the entities were unknown to all readers because they were represented by nonce (nonsense) words. In contrast, two of the entities referred to were likely to be known to all children. In addition, the narrative compared or contrasted the entities along a dimension that was likely to be familiar to all children, and the known entities differed strikingly on that dimension. The narrative text and associated questions were divided into three sections presented successively, in an effort to minimize the memory demands of the task for young children. Each new section consisted of roughly two propositions and the associated questions. After reading each section, students were asked a series of yes-no questions about that section. Students could reread the previously presented material followed by the new section before answering the questions associated with the second and third sections. Successive sections also included questions that related propositions across previously presented sections

of the text. As in the Potts and Peterson task, the questions were framed to assess students' abilities in each of the four components of reading comprehension described above. We developed two passages: "Nan's Pets" and "Tom and Ren." Both passages were developed first in English and then translated into Spanish using back translation. In study 1 either "Nan's Pets" or "Tom and Ren" was administered in English only. In study 2 "Nan's Pets" was administered in English and "Tom and Ren" was administered in Spanish. In study 3 children received both an English and a Spanish passage, with assignment of passage to language determined randomly but with the provision that no child received the same passage in both English and Spanish. A copy of the English version of "Nan's Pets" can be found in the appendix.

Procedures

Prior to administering the assessment, we gave children a practice story (see App.). This story served three purposes. Students read the story aloud and were rated for accuracy of word reading; children who scored below 85% on word accuracy were not administered the assessment. Second, it provided all children an opportunity to practice answering the kinds of questions they were to be asked during the assessment and provided a chance for the examiner to explain why students' answers were correct or incorrect. Because the text included at least some entities that had been named with nonsense words, the practice text also prepared students to read texts where some words would be unknown to them

The participants in the first and second pilot studies were administered the assessment individually in a clinical interview mode; children were asked to justify their answers, and the tester probed to be sure she understood the children's reasoning for each answer. The responses were tape-recorded, transcribed, and analyzed qualitatively. The subjects in the second sample were given slightly revised versions of the assessment. Because we had available the participants' justifications for their answers, we could analyze their responses in two ways: simple responses (i.e., one point for correct, and no points for incorrect, answers, totaled for each of the components), and justified responses. We calculated justified responses by assigning 0 for no response or an incoherent or incorrect justification, 1 for a plausible or possibly correct but incomplete explanation, and 2 for a full and correct justification. The simple response score was then multiplied by the justification code to generate a justified score. Subjects in the third study, in addition to being tested on the DARC, were assessed on word reading, word reading efficiency, comprehension, and language proficiency using the Woodcock Language Proficiency Battery (WLPB; Woodcock, 1991).

Results

The three pilot studies indicated that it is possible to use the DARC with students as young as kindergarten, to develop texts that are easily decoded but yet place demands on students' abilities to form inferences and integrate information across propositions and with background knowledge, that the yes/no scores accurately reflect students' elaborated responses, and that the test can be used to differentiate inferencing from text memory and background knowledge.

Pilot Testing Items

Despite the cognitive demands of the assessment, children at all grade levels (in the Washington, DC, sample) were able to complete the assessment (see Table 1). Furthermore, students who scored quite differently on the Stanford-9 scored similarly on the DARC (e.g., see the second graders), and students who scored identically on the Stanford-9 got very different results on the DARC (see the sixth graders). These preliminary results confirmed that the DARC was tapping comprehension capacities somewhat different from those measured on the

TABLE 1. DARC Total and Component Scores and Stanford-9 Reading Scores for Children in Pilot Study 1 ($N = 16$)

	DARC Score					Standford-9 Spring 2001 (NCE)	
Grade/ID	Total	Text Memory	Text Inferencing	Background	Knowledge Integration	Total Reading	Reading Comprehension
2/1	20	6	2	5	7	68.5	64.2
2/2	23	9	2	6	6	69.3	71.8
2/3	22	7	1	7	7	40.7	45.2
3/1	27	10	3	7	7	74.7	82.7
3/2	16	4	1	6	5	49.5	49.5
3/3	24	8	2	7	7	59.8	62.9
3/4	21	6	1	6	8	57.5	56.4
4/1	28	6	5	7	10	56.4	57.0
4/2	27	11	2	7	7	51.1	52.1
5/1	28	6	5	5	12	40.2	42.5
5/2	21	4	3	7	7	52.1	51.1
5/3	28	10	3	6	9	55.9	55.9
5/4	26	9	2	6	9	71.8	79.6
6/1	30	6	5	7	12	59.3	59.3
6/2	24	6	4	4	10	*na*	*na*
6/3	24	10	3	5	6	57.5	55.3
Maximum score:							
Story 1	30	6	5	7	12		
Story 2	30	11	3	7	9		

NOTE.—NCE = normal curve equivalent.

standardized reading comprehension assessment.

Examination of responses to the clinical interview suggested that some sentence constructions used in the statements to be judged true or false confused the participants. In subsequent versions of the test, we modified these constructions. Specifically, the items to test knowledge integration were simplified by eliminating the leading clause "Like XXX." For example, in "Nan's Pets," the item "Like crabs, snerps have shells to protect them" was changed to read "Snerps have shells to protect them" (see App.). The objective of the item was to have students integrate background knowledge (turtles have shells) with text knowledge (snerps are like turtles) and arrive at the correct answer: snerps are like turtles and turtles have shells, so snerps have shells too (six items from "Nan's Pets"). In addition, one background knowledge item (#18) was replaced with a knowledge integration item, and one knowledge access item (#19) was changed to ensure that we tested background knowledge essential to text inferencing; specifically, we deleted "Culps have fur to keep them warm" and replaced it with "A turtle is faster than a dog." With regard to scoring, we changed one item from text memory to text inferencing ("Nan has a dog") because children had to infer that this was incorrect (the story indicated that Nan had a cat). We made similar revisions to the story entitled "Tom and Ren."

Using the DARC with Poor Overall Comprehenders

In the second pilot study (in Chicago and Boston) we tested participants who had displayed very poor performance on the reading comprehension subtest of the Woodcock Language Proficiency Battery (WLPB) administered in English. Thus, we were primarily interested in knowing whether these children in general, or some subset of them, performed well on the DARC. Good performance on the DARC would suggest that they had adequate text memory, background knowledge, and inferencing abilities but were unable to use those abilities when challenged by text that was difficult to decode, grammatically complex, or filled with unfamiliar vocabulary items. Furthermore, we tested these children on the Spanish version of the DARC to collect additional evidence about whether their basic comprehension skills were intact. Finally, we were interested in examining the relative difficulty of the four subscales and exploring whether children's yes/no responses were produced by random guessing or reflected their reasoned analysis and/or memory of the text.

Figure 1 presents the distribution of participants' scores on the DARC plotted against their scores on the WLPB passage comprehension measure; it is clear that some children performed well on the DARC in English despite having low scores on the standardized comprehension measure. These results are not simply due to the DARC being easier than the WLPB passage comprehension, because it is clear that some students who performed relatively well on the WLPB scored rather poorly (i.e., near chance levels) on the DARC. These relations merit reexamination in the larger sample.

Second, what was the overall performance of these bilingual children on the DARC, and was it better in English or Spanish? For the 25 children who could be tested in both languages (three refused to take the test in Spanish), there was no significant difference between the two languages. Students scored 20.16 on average in English (out of a maximum of 30), and 19.40 in Spanish (see Fig. 2). We are, of course, assuming in presenting this comparison that the items are of equal difficulty across the two languages; although we designed the test with formally equivalent items in English and Spanish, we have not, with this small sample, carried out the psychometric analysis needed to demonstrate that the two versions are fully equivalent. Furthermore, although all these participants were native Spanish speakers who had received initial

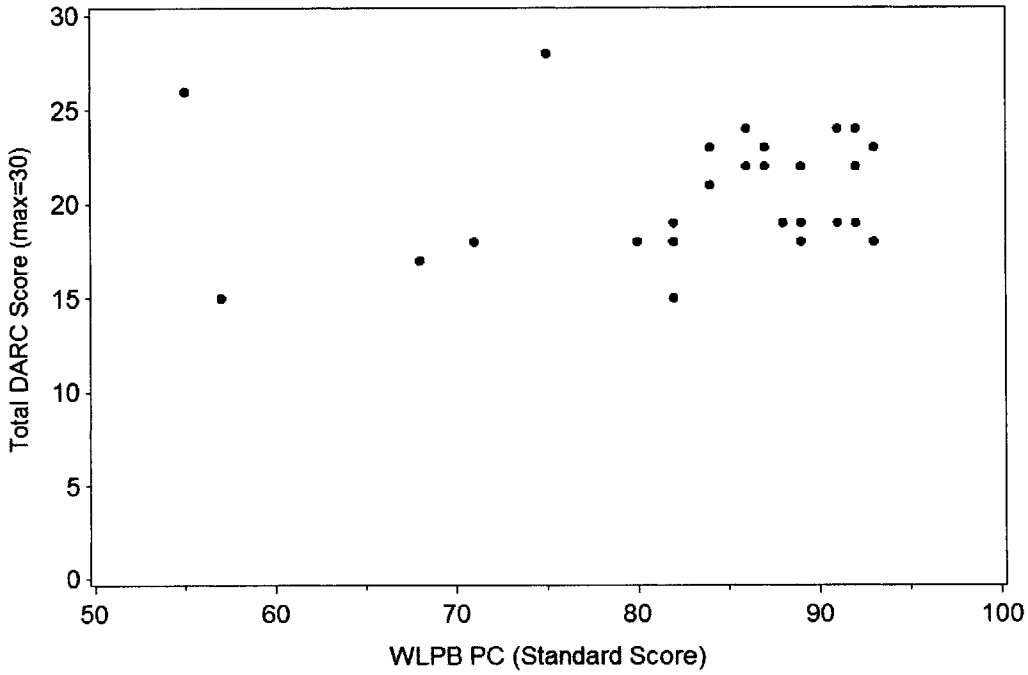

FIG. 1.—Relation between DARC total score and WLPB passage comprehension (English—standard score) for children selected because of low performance on the WLPB ($n = 28$).

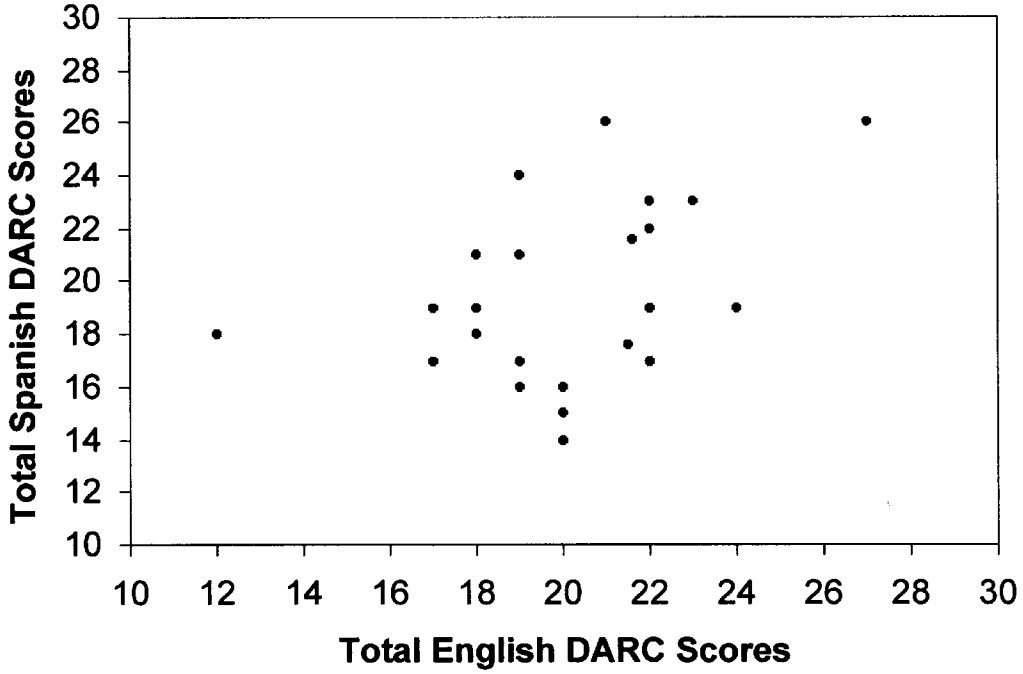

FIG. 2.—Correlation between Spanish and English DARC scores for the Boston/Chicago sample ($n = 28$)

literacy instruction in Spanish, their Spanish proficiency at the time of this testing was not independently assessed and may well have influenced these relations.

Six of the 25 bilingual students achieved exactly the same score in English and Spanish, 11 scored higher in English, and eight scored higher in Spanish. All three subgroups showed a wide range of scores; in other words, there were relatively poor readers and relatively good readers in all three groups. The correlation between Spanish and English total scores was .39 ($p < .05$), significant but in the low-moderate range. That, on average, children scored at comparable levels in Spanish and English constitutes preliminary evidence that the DARC decreased the English-language demands of the English assessment while maintaining the cognitive challenge.

The DARC was intended to provide scores on four components of comprehension—text memory, text inferencing, background knowledge, and knowledge integration. It stands to reason that knowledge integration and text inferencing would be more difficult for young children than text memory and background knowledge. For the most part, that is the pattern of difficulty observed in the participants' performance. In English, students performed best, in order of percentage correct, on text memory, then knowledge access, knowledge integration, and text inferencing. In Spanish, knowledge access was easiest, then text memory, knowledge integration, and text inferencing (see Table 2). The finding that knowledge integration was only slightly more difficult than background knowledge in both English and Spanish may indicate that knowledge integration items relied too heavily on known attributes of the entities in the stories. This relation merits further examination in larger samples and possibly continued work on the development of knowledge integration items. It is also worthy of mention that performance on the text inferencing subscale was not different from chance in either English or Spanish; there were also very few items on this subscale, a feature that clearly needs to be improved in future versions of the DARC.

Importantly, the four components of the DARC were relatively independent of one another, as would be predicted by the theory underlying the construction of the test. Scores were not completely unrelated, but bivariate correlations between subscales (see Table 3) as well as correlations between each of the individual subscales and the other three subscales (presented on the diagonal in Table 3) ranged from small to moderate; only the correlation between text inferencing and background knowledge was significant at $p < .05$. Based on this preliminary sample of cases, we concluded that the four subscales provided relatively nonoverlapping information regarding reading comprehension as predicted. Of course, the relation among subscales needs to be reexamined in a larger sample unselected for performance on an external comprehension measure, but the results are at least promising that the different subscales assess different information about comprehension. This conclusion is further supported by the intersubscale correlations presented in Table 3.

TABLE 2. Mean Number and Percentage Correct for DARC Subscales, in English and Spanish, Pilot Study 2

	English ("Nan's Pets")				Spanish ("Tom and Ren")			
Subscale	No. of Items	Mean	*SD*	%	No. of Items	Mean	*SD*	%
Text memory	6	5.08	.76	85	11	7.88	1.54	72
Text inferencing	5	2.28	1.46	46	3	1.56	.87	52
Background knowledge	6	4.76	1.16	79	5	4.24	.88	85
Knowledge integration	13	8.04	1.65	62	11	5.80	2.16	53

TABLE 3. Correlations among DARC Subscales and Corrected Subscale-Total Correlations[a] (on Diagonal), Pilot Study 2

Subscale	Text Memory	Text Inferencing	Background Knowledge	Knowledge Integration
Text memory	.17			
Text inferencing	.30	.35^{+}		
Background knowledge	−.04	.37*	.28	
Knowledge integration	.06	.09	.12	.13

[a]Correlation between component score and total for remaining components (i.e., total excluding items from the component being correlated).

$^{+}p < .10$.

*$p < .05$.

An important component of the test development process was the evaluation of the basis for students' responses to the true-false questions. Of course, students can answer questions correctly simply by guessing, because the questions required only yes or no in response. Consequently, we wanted to ascertain the extent to which students had correctly reasoned about the relations among entities in the test, and the extent to which students' dichotomized responses reflected their underlying thinking about the text and the relation among entities. To address this question, we examined students' responses to the follow-up questions and scored these as correct or incorrect. We refer to these scores as justified scores because students had to justify their answer after giving it. These justified scores were then compared to students' original answers, which we refer to below as simple scores.

In general, results supported the utility of the simple scoring procedure. Specifically, in English and in Spanish, the correlation between the simple and the justified scores was .91, which was statistically significant at the $p < .01$ level. Moreover, correlations between WLPB passage comprehension scores and scores for the four subscales showed that relations with the simple and the justified scores were not appreciably different from one another (see Table 4). Thus, students' simple responses reflected their thinking about the passages and about the true-false items.

Large-Sample Validation Study

In this large-scale Texas validation study of the DARC with Spanish-English bilingual students in kindergarten through grade 3, 521 participants (K = 12; grade 1 = 130; grade 2 = 180; grade 3 = 198) took the DARC in English and Spanish, along with the passage comprehension and other subtests from the Woodcock Battery (WLPB) in both English and Spanish. This sample included only students who were able to read the passages on the DARC in English using one of the two forms; a few students who took the English form were unable to read the Spanish version. There was an orderly pattern of increase with grade in number of items correct (see Table 5) on both stories used, in both Spanish and English. At the same time, there was no evidence of a ceiling effect even among third graders.

Internal consistency reliability (Cronbach's alpha) for the subscales ranged between .41 and .54 in English and from .21

TABLE 4. Spearman Rank Correlations of DARC with WLPB English Passage Comprehension, Pilot Study 2

DARC Subscale	Justified Score	Simple Score
Total score	.09	.08
Text memory	−.12	−.08
Text inferencing	−.06	−.02
Background knowledge	.11	.34[a]
Knowledge integration	−.06	−.15

[a]Significant at the $p < .10$ level. All other correlations are not significantly different from 0.

TABLE 5. Mean Total Number Correct Responses (of 30 Possible) for Grades K–3 on Stories 1 and 2 in English and Spanish, Pilot Study 3

	Story 1			Story 2		
Grade/Language	*N*	Mean	*SD*	*N*	Mean	*SD*
Kindergarten:						
English	7	22.43	2.94	5	19.40	2.51
Spanish	16	18.56	18.56	19	17.74	3.05
Grade 1:						
English	64	21.25	4.69	67	20.12	4.30
Spanish	68	19.85	4.06	68	19.96	3.47
Grade 2:						
English	97	21.95	3.96	83	22.35	3.74
Spanish	71	21.97	3.92	78	21.08	3.35
Grade 3:						
English	105	23.49	4.15	93	23.12	3.73
Spanish	88	22.73	3.82	88	22.54	3.84

TABLE 6. Correlations among Subscales, Corrected Subscale-Total Correlations[a] (on Diagonal), and Disattenuated Correlations[b] (above Diagonal), Pilot Study 3

Language/Subscale	Text Memory	Text Inferencing	Background Knowledge	Knowledge Integration
English (n = 521):				
Text memory	.60	.39	.67	.24
Text inferencing	.17	.63	.81	.75
Background knowledge	.32	.39	.74	.76
Knowledge integration	.11	.34	.38	.77
Cronbach's alpha: Form 1	.41	.45	.54	.47
Form 2	.46	.43	.52	.47
Spanish (n = 496):				
Text memory	.54	.86	.89	1.14
Text inferencing	.32	.58	.78	1.04
Background knowledge	.35	.30	.59	1.13
Knowledge integration	.43	.38	.44	.66
Cronbach's alpha: Form 1	.21	.43	.36	.46
Form 2	.50	.28	.45	.25

NOTE.—Alpha for TI and KI Combined: Form 1—English = .64, Spanish = .62; Form 2—English = .52 (.59 with item 10 deleted), Spanish = .34 (.44 with item 10 deleted); total scale alpha: Form 1—English = .75, Spanish = .67; Form 2—English = .70, Spanish = .64.

[a]Correlation between component score and total for remaining components (i.e., total excluding items from the component being correlated).

[b]Disattenuated correlations above the diagonal are equal to the observed correlation divided by the square root of the product of the reliabilities. To estimate scale reliabilities, we took the square root of the average of the squared reliability for Form 1 and Form 2 in a given language.

to .50 in Spanish for each subscale and form, and from .64 to .75 for the total score (see Table 6). It is worth pointing out that reliabilities for each scale were satisfactory (.4 or above) given the current use of only a single text to elicit student responses. However, the same cannot be said for each scale-form-language combination. The most problematic of these combinations were the two inferencing scales on Form 2 in Spanish. The discrepant internal consistencies within Spanish across the two forms and between languages on Form 2 indicated that more work is needed to build forms that are truly comparable, that is, parallel or interchangeable across and within languages. Because individual scale reliabilities were not high for either form or language, we also tried

combining the inferencing scales (text inferencing and knowledge integration) into a single scale. The combined inferencing scale yielded alphas of .64 and .62 for Form 1 in English and Spanish, respectively, and .59 and .44 for Form 2 with item 10 deleted in English and Spanish, respectively. Item statistics indicated that item 10 on Form 2 was problematic as an inferencing item. Thus, although scale reliabilities were adequate for Form 1 in both languages and for Form 2 as well in English, the Spanish adaptation of Form 2 requires additional work.

In addition to scale reliabilities in each language for each form and scale, Table 6 presents correlations for each language, collapsing across forms. For the correlations presented below, scores were standardized for each form and then combined to give estimates of English and Spanish reading ability as measured on the DARC. In addition, Table 6 presents corrected correlations between each scale and the total score on the DARC, excluding the scale being correlated. These correlations appear on the diagonal of Table 6.

The subscale correlations with the total DARC score ranged from .60 to .74 in English, and from .54 to .66 in Spanish. In both English and Spanish, intercorrelations among subscales were considerably lower, ranging from .17 to .39 in English and from .32 to .44 in Spanish. Thus, the pattern of relative independence among the subscales found in the Boston/Chicago pilot study was replicated with this much larger sample. Above the diagonal in Table 6, we provide correlations among the subscales, disattenuated for unreliability (Kenny, 1979). For the English version of the test, these disattenuated correlations show that the inferencing scales were only somewhat related to text memory and were more highly related to one another and to background knowledge. The disattenuated correlations in Spanish showed the scales to be less differentiated. That some of the disattenuated correlations exceeded 1.0 indicated that the reliability of the scales was likely underestimated by Cronbach's alpha and shows that the Spanish adaptation requires additional work.

Table 7 provides correlations of the DARC with subtests from the Woodcock Language Proficiency Battery (WLPB). The DARC was less highly correlated with measures of decoding ($r = .28$ and .22 for WLPB letter word and word attack, respectively) than the WLPB passage comprehension ($r = .65$ and .62, respectively). In other words, performance on the WLPB passage comprehension test was much more influenced by decoding skills than was performance on the DARC. The DARC showed much stronger correlations with listening comprehension and oral language from the WLPB (.46 and .53, respectively) than it did with decoding, as one would hope and expect. In contrast, the WLPB passage comprehension correlated .64 and .72 with listening comprehension and oral language, that is, about as highly as it correlated with the two decoding measures. (All correlations were significant at $p < .0001$.) Thus, despite the simplicity of the passages presented in the DARC, performance was strongly related to oral language processing, as reading comprehension should be, and less affected by word-reading skills than the WLPB. More complete information on correlations of the DARC total and subscale scores with reading and language measures from the WLPB is presented in Table 7. In this table, all correlations are significant at $p < .0001$ with the exception of the correlation between WLPB letter word and text memory, which is not statistically significant. Thus, data from this robust validation study indicated that our first efforts at developing the DARC were successful in making test performance more dependent on higher-order processing and less dependent on word-level decoding skills.

The correlations in Table 7 show that, of the four DARC subscale scores, background knowledge correlated most highly with the WLPB passage comprehension score. More

important, WLPB passage comprehension appeared more highly related to letter-word identification, a measure of decoding skill (.65), than any of the four DARC subscales (.05 to .29). Finally, all the DARC subscale scores were moderately to strongly correlated only with measures of oral language and listening comprehension, whereas the WLPB passage comprehension correlated with decoding as strongly as with these measures.

In Table 8 we present correlations between English and Spanish for the DARC total scores and measures of reading and language proficiency taken from the WLPB. These correlations are included to show the extent of intra- and interlanguage correlation for the DARC and WLPB. The correlations again show that the DARC was less correlated with word reading than was the WLPB passage comprehension subtest in both English and in Spanish. Interlanguage correlations were different for the DARC and WLPB measures. The DARC in Spanish correlated positively with the DARC in English as well as with English measures of word reading, reading comprehension, and oral language proficiency. In contrast, the DARC in English correlated negatively with reading comprehension and word reading in Spanish and negligibly with oral language proficiency in Spanish. Furthermore, reading comprehension and word reading in Spanish measured by the WLPB correlated negatively with oral language proficiency in English and reading comprehension measured by the DARC but not by the WLPB. Interestingly, the interlanguage correlations for comprehension measured on the DARC and the WLPB were virtually

TABLE 7. Pearson Correlations for English Measures of Reading and Language with English scores from the WLPB Comprehension and the DARC, for the Texas Sample ($n = 521$)

		DARC				
WLPB Scale	WLPB Passage Comprehension	Total	Text Inferencing	Text Memory	Background Knowledge	Knowledge Integration
Passage comprehension		.28	.18	.12	.31	.20
Listening comprehension	.64	.46	.31	.22	.48	.29
Oral language	.72	.53	.38	.24	.55	.34
Letter-word identification	.65	.28	.24	.05	.29	.21
Word-reading efficiency	.32	.34	.19	.16	.38	.25

TABLE 8. Pearson Correlations between English and Spanish Reading and Language Measures for Children Who Read in Both Languages on the DARC ($n = 366$), Pilot Study 3

	English			Spanish		
		WLPB			WLPB	
Measure	DARC	Passage Comprehension	Letter-Word	DARC	Passage Comprehension	Letter-Word
English (WLPB):						
Passage comprehension	.32					
Letter-word	.35	.67				
Oral language	.57	.73	.65	.26	−.07	−.23
Spanish:						
DARC	.28	.14	.20			
WLPB:						
Passage comprehension	−.14	.28	.20	.17		
Letter-word	−.20	.03	.16	.10	.71	
Oral language	.06	.08	.17	.37	.68	.56

NOTE.—Correlations larger than .13 in absolute value are significant at $p < .01$.

identical ($r = .28$, $p < .0001$) despite their differing patterns of correlations with word reading and oral language.

Although these data are only preliminary, they offer promising indications that it may be possible to build a measure of comprehension that is not heavily influenced by decoding skill but that remains sensitive to the language and thinking skills of ELLs.

Discussion

We have presented findings from three pilot studies that demonstrate the potential value, usability, and discriminative capacity of a new diagnostic assessment of reading comprehension. These preliminary analyses show that the DARC is feasible for use with children as young as kindergartners, that simple yes-no responses reflect children's comprehension processing on the DARC adequately, that different aspects of the comprehension process (text memory, text inferencing, background knowledge, and knowledge integration) can be measured independently, and, most important, that this measure reveals children's comprehension capacities that are obscured by measures with greater decoding, syntax, and vocabulary load. Some children who score poorly on the Stanford-9 or the WLPB passage comprehension measure perform well on the DARC, suggesting that their poor performance on the standardized measure reflects difficulties with some part of the comprehension process (e.g., word decoding, vocabulary) other than comprehension processing per se. And the DARC is much less influenced by differences in word-reading skills than is the standardized passage comprehension subtest of the WLPB.

Thus, the DARC design has strong advantages over more traditionally designed, portmanteau comprehension measures and may be particularly useful for assessing comprehension processing among English-language learners and other groups of children with limited vocabulary. Standard instruction for learners scoring low on comprehension assessments is likely to focus on providing them with strategies for improving comprehension; students who score poorly on general comprehension tests but well on the text inferencing and knowledge integration subscales of the DARC probably do not need such instruction. Instead, their control over the language demands of the texts they are reading and their access to relevant background knowledge are more likely explanations for their comprehension problems and suggest a very different focus for intervening with them.

Findings from the third pilot study, the validation study carried out in Texas, indicate that additional work is necessary to improve the reliability of scores on the text inferencing, knowledge integration, text memory, and background knowledge subscales. Subscale reliabilities in English range from .41 to .54, as compared to the full test reliabilities of between .70 and .74. By creating additional passages, we can increase the number of items measuring each of these components while reducing dependence among items. Increasing subscale reliabilities is clearly a prerequisite to developing the DARC into a useful diagnostic instrument.

Additional pilot work is also needed to evaluate the DARC with larger groups of fourth and fifth graders, to determine whether restricting the passages to a second-grade reading level creates problems in assessing older children and to consider the effect of including children with somewhat greater word-reading difficulties in the sample.

The preliminary evidence presented here offers considerable promise that the DARC can be further developed, expanded, and extended in range. Although we have focused in this article on its value for Spanish speakers learning English, the test is also of potential value in helping to pinpoint sources of comprehension difficulties for English-only students who score poorly on more general measures. Precisely because comprehension is such a multidetermined process, knowing how to go about instruct-

ing and intervening most efficiently offers a great challenge. Traditional comprehension assessments offer classroom teachers little guidance about the needed instructional focus. The DARC, building on the work of Potts and Petersen (1985) and Hannon and Daneman (2001), offers the potential to improve and individualize comprehension instruction by providing teachers with subscale scores that reflect the specific components of reading comprehension with which students, including English-language learners, might be having difficulties.

The DARC also has great potential as a research instrument because it offers the possibility of varying characteristics of the texts being read. For example, although the passages we used in these pilot studies were designed to be simple in vocabulary load and syntactic structure, it would be possible to develop additional, more difficult items in the areas of vocabulary knowledge, syntactic complexity, or presupposed background knowledge. If a learner who scored well on the central comprehension processes in passages like "Nan's Pets" scored poorly on one of these additional passages, that would suggest alternative sources of comprehension difficulties. Development of such text manipulations is not outside the realm of feasibility but has not yet been attempted. Nevertheless, even without these additions, the preliminary work to date indicates that a more fully developed standard set of forms for the DARC would provide educators and researchers with a tool for assessing young students' performance on components of comprehension that are not easily extracted from scores on existing comprehension tests in use in the elementary grades.

Appendix

Diagnostic Assessment of Reading Comprehension (Pilot Version)

PRACTICE TEXT

Maria likes to eat fruit. Most of all she likes to eat orkers. An orker is like an orange. But an orker is bigger than an orange.

PRACTICE ITEMS

1) Maria likes to eat fruit.
   - If correct: That's right. The story tells us Maria likes to eat fruit.
   - If incorrect: "How do you know that?" or "What does the story tell us?"
2) Most of all Maria likes to eat orkers.
   - If correct: That's right. The story tells us Maria likes to eat orkers.
   - If incorrect: Look at the story again. [Represent the practice story stimulus.] What does the story tell us that Maria likes to eat most of all?
3) An orange has a peel.
   - If correct: That's right. The story does not tell you that an orange has a peel, but you know this from your everyday life. An orange has a peel or a skin that you take off before you eat the orange.
   - If incorrect: The skin of an orange is called a peel. Even though the story does not say that an orange has a peel, you know that an orange has a peel or skin from your everyday life. So, you would answer this question, "Yes. An orange has a peel."
4) You peel an orker to eat it.
   - If correct: That's right. An orker is like an orange and you peel an orange to eat an orange. That's how we know that you peel an orker to eat an orker.
   - If incorrect: Listen to the story again. [Repeat story.] What does the story tell us about oranges and orkers? The story tells us that an orker is like an orange. Do you peel an orange? Yes, that's right, you peel an orange to eat it. If an orker is exactly like an orange, do you think you peel an orker to eat it? That's right, you peel an orker to eat it.

STORY TEXT

Nan has four pets. One pet is a cat. Nan's cat is fast. Nan has a pet culp. Nan's pet culp is like her cat. But Nan's pet culp is faster than her cat.

Nan has a pet turtle. Nan's turtle is slow, Nan also has a pet tarf. Nan's pet tarf is like her pet turtle. But Nan's pet tarf is slower than her turtle.

One day Nan got a pet snerp. Now Nan has five pets. Nan's snerp is like her tarf. But Nan's snerp is slower than her tarf. All of Nan's pets like to play. The pets like to play in Nan's backyard.

STORY ITEMS (true-false)

1) Nan has four pets.
2) Nan's cat is slow.
3) Cats have fur to keep them warm.
4) Nan's pet cat is faster than her pet culp.
5) Culps have fur to keep them warm.
6) Nan has a turtle.
7) Nan's tarf is like her culp.
8) Nan's turtle is faster than her tarf.
9) A cat is faster than a tarf.
10) A culp is faster than a turtle.
11) A tarf is faster than a dog.
12) Tarfs have fur to keep them warm.
13) Nan got a pet snerp.
14) Nan has a dog.
15) A turtle is slower than a cat.
16) Cats can live in water.
17) Turtles have fur to keep them warm.
18) A culp is faster than a crab.
19) A turtle is faster than a dog.
20) Turtles cannot live in water.
21) Now Nan has five pets.
22) Nan's snerp is like her culp.
23) Nan's pets like to play in her backyard.
24) A culp is faster than a snerp.
25) The turtle is slower than the snerp.
26) A snerp is faster than a cat.
27) Snerps have shells to protect them.
28) Snerps live in trees.
29) Snerps can live in water.
30) Nan's tarf is slower than her snerp.

Note

This research was supported in part by grants PO1 HD039530, "Acquiring literacy in English: Cross-linguistic, intra-linguistic, and developmental factors," and PO1 HD39521, "Oracy/literacy development of Spanish-speaking children," both jointly funded by the National Institute of Child Health and Human Development of the National Institutes of Health and the Institute of Education Sciences of the U.S. Department of Education. The opinions expressed herein are ours and do not necessarily reflect the opinions of, or endorsement by, the funding agencies.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read: A causal connection. *Nature,* **303,** 419–421.

Dixon, P., LeFevre, J., & Twilley, L. C. (1988). Word knowledge and working memory as predictors of reading skill. *Journal of Educational Psychology*, **80,** 465–472.

Engle, R. W., Nations, J. K., & Cantor, J. (1990). Is "working memory capacity" just another name for word knowledge? *Journal of Educational Psychology*, **82,** 799–804.

Gathercole, S. E., & Pickering, S. J. (2000). Working memory deficits in children with low achievement in the national curriculum at seven years of age. *British Journal of Educational Psychology*, **70,** 177–194.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education,* **7**(1), 6–10.

Guthrie, J. T., Wigfield, A., & VonSecker, C. (2000). Effects of integrated instruction on motivation and strategy use in reading. *Journal of Educational Psychology,* **92**(2), 331–341.

Haengi, D., & Perfetti, C. A. (1994). Processing components of college-level reading comprehension. *Discourse Processes*, **17,** 83–104.

Hannon, B., & Daneman, M. (2001). A new tool for understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology,* **93**(1), 103–128.

Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday lives of young American children*. Baltimore: Brookes.

Hulme, C., Muter, V., Snowling, M., & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: Evidence from a longitudinal study. *Developmental Psychology,* **40,** 665–681.

Kenny, D. (1979). *Correlation and causality*. New York: Wiley.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* Washington, DC: National Institute of Child Health and Human Development

National Research Council. (1997). *Improving schooling for language-minority children.* Washington, DC: National Academies Press.

National Research Council. (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academies Press.

Palmer, J., MacLeod, C. M., Hunt, E., & Davidson, J. E. (1985). Information processing correlates of reading. *Journal of Memory and Language*, **24,** 59–88.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford.

Potts, G. R., & Peterson, S. B. (1985). Incorporation versus compartmentalization in memory for discourse. *Journal of Memory and Language*, **24,** 107–118.

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Washington, DC: RAND Education.

Stanovich, K. E. (1991). Word recognition: Changing perspectives. In R. Barr, M. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. **2**, pp. 418–452). New York: Longman.

Sweet, A. P., Guthrie, J. T., & Ng, M. M. (1998). Teacher perceptions and student motivation. *Journal of Educational Psychology*, **90**(2), 210–223.

Vellutino, F. R. (1979). *Dyslexia: Theory and research*. Cambridge, MA: MIT Press.

Vellutino, F. R. (1987). Dyslexia. *Scientific American,* March, pp. 34–41.

Woodcock, R. W. (1991). *Woodcock Language Proficiency Battery—Revised (English Form)*. Chicago: Riverside.